

支持鹏程系列开源大模型应用生态演化的可持续学习能力探索

余跃^{1,2}, 刘欣¹, 蒋芳清¹, 张晗¹, 王晖¹, 曾炜³

(1. 鹏城实验室网络智能部开源所, 广东 深圳 518055;

2. 国防科技大学, 湖南 长沙 410073;

3. 北京大学, 北京 100091)

摘要: 大规模预训练模型利用大规模语料以及多样化的预训练任务在自然语言处理领域取得了巨大的成功。随着大模型的逐步发展, 大模型的可持续学习能力探索成为新的研究热点。主要介绍鹏程系列大模型持续学习的技术体系、应用实践以及面临的挑战, 包括通过任务扩展、数据增量和知识推理的鹏程系列可持续学习技术体系, 开源大模型鹏程·盘古多任务可持续学习和鹏程·通言大模型的可持续学习能力实践, 大模型可持续学习过程中面临的词表更新、语义映射和知识冲突等挑战。

关键词: 鹏程系列大模型; 可持续学习; 鹏程·盘古; 鹏程·通言; 开源大模型

中图分类号: TP391.1

文献标志码: A

doi: 10.11959/j.issn.2096-6652.202212

Exploration of the continual learning ability that supports the application ecological evolution of the large-scale pretraining Peng Cheng series open source models

YU Yue^{1,2}, LIU Xin¹, JIANG Fangqing¹, ZHANG Han¹, WANG Hui¹, ZENG Wei³

1. Open Source Institution, Network Intelligence Department, Peng Cheng Laboratory, Shenzhen 518055, China

2. National University of Defense Technology, Changsha 410073, China

3. Peking University, Beijing 100091, China

Abstract: Large-scale pre-training models have achieved great success in the field of natural language processing by using large-scale corpora and pre-training tasks. With the gradual development of large models, the continual learning ability of large models has become a new research focus. The continual learning technology of the Peng Cheng series large models, the exploration of practice and the still facing challenges were mainly introduced, including the Peng Cheng series continual learning technology through task expansion, data increment and knowledge reasoning, Peng Cheng PANGU multi-task continual learning and the practical exploration of the continual learning ability of the Peng Cheng TONGYAN open source large model, the vocabulary update, semantic mapping and knowledge conflicts that the large model faces in the process of continual learning.

Key words: Peng Cheng series large model, continual learning, Peng Cheng PANGU, Peng Cheng TONGYAN, open source large model

0 引言

近年来, 鹏程系列大规模预训练语言模型(简

称鹏程系统大模型)通过自监督学习的训练方式从海量语料中学习文本的上下文表示, 并通过针对特定任务的小样本学习, 在自然语言理解和自然语言

收稿日期: 2021-12-01; 修回日期: 2022-01-13

通信作者: 王晖, wangh06@pcl.ac.cn

基金项目: 新形势下我国技术开源战略研究(NO.GHZX2020ZCQ013)

Foundation Item: Research on Chinese Technological Open Source Strategy Under New Situation (No.GHZX2020ZCQ013)

生成等任务上都取得了惊艳的成果^[1-2]。

大模型具有两个显著的特点，一是模型结构足够深且参数规模庞大，二是预训练过程中伴随着多种预训练任务以保证大规模语料的有效利用。随着谷歌发布具有1~3亿参数规模的预训练的BERT模型^[3]，大规模预训练模型正式进入自然语言处理领域并开始蓬勃发展。随后几年，国内外研究机构发布的预训练模型的参数规模不断增长，依次出现了XLM（约2亿）^[4]、RoBERTa（约3.5亿）^[5]、MT5（约7亿）^[6]、GPT-3（约1750亿）^[7]、华为盘古大模型（约1000亿）、鹏程·盘古- α 大模型（约2000亿）^[2]、谷歌switch transformers（约1.6万亿）^[8]、智源悟道2.0（约1.75万亿）以及阿里达摩院M6（约1万亿）^[9]等大规模预训练模型。在预训练大模型中，预训练任务占据着举足轻重的地位，主要包括概率化语言模型^[10]、完形填空式掩码语言模型及其变体^[3]、序列到序列掩码模型^[11]、排列语言模型^[12]、降噪自编码模型^[12]和句子预测及连贯性等，多样化的预训练任务为较深结构模型从大规模语料中学习丰富的语言知识提供了有效的途径。在近两年发布的大模型中，鹏程·盘古- α 大模型、谷歌switch transformers、智源悟道等大模型都选择了开源开放发展策略，进而推动相应技术的应用落地与生态建设。

然而，大模型中的参数规模和预训练任务这两个特点也决定了训练好的大模型很难进行新数据、新结构的学习，这主要体现在两个方面。一是模型训练资源消耗较大，训练一个大模型通常需要较多硬件资源和时间，因此面对新数据时重新训练模型会浪费较多资源；二是模型结构复杂，模型结构的变动会导致原有预训练好的参数无法在新结构中适用，重新训练又会重复资源浪费。因此，在大模型上开展与可持续学习相关的研究以推进大模型的可持续演化是十分必要的研究方向。尤其是在大模型开源化发展的趋势下，本项研究对于以开源带动群体创新，持续构建大模型开源生态体系具有重要意义。

可持续学习是一个连续不断的学习过程，在这个过程中，模型已经完成了 N 个学习任务($T_1, T_2, \dots, T_N$)。当面对第 $(N+1)$ 个任务 T_{N+1} 和它的数据 D_{N+1} 时，模型可以利用其知识库中的先验知识来帮助学习 T_{N+1} ^[13]。知识库存储并维护在过去学习 N 个任务时所学习和积累的知识。在学习了 T_{N+1} 之后，知识

库更新了从 T_{N+1} 中学习到的结果^[13]。在大模型上应用持续学习后，模型应该具有动态接收新数据、新结构的能力，而不需要通过重新训练这样的高资源消耗方法来实现知识的学习。因此，经过成功的持续学习的模型应该具有知识迁移的能力：任务或者数据间的相关性使得模型在学习完相似的任务或者数据时，各自的性能相较模型学习单任务时都有所提升且语言信息更加丰富。通过可持续学习演进，模型将具有较强的可塑性（学习新知识的能力）和稳定性（旧知识的记忆能力）^[14]，同时可以解决大模型传统演进过程中面临的资源消耗问题，因此可持续学习能力对于大模型应用生态的持续演化来说是至关重要的。

本文组织结构如下：第1节介绍可持续学习相关研究，包括可持续学习方法概述和鹏程系列大模型的可持续学习能力技术体系；第2节介绍鹏程·盘古多任务可持续学习的实践探索，包括基于预训练大模型的多任务持续学习方法、相关实验设置以及实验结果；第3节介绍鹏程·通言大模型的可持续学习能力探索，包括参数结构的增量式学习和翻译语言的可持续学习；第4节对鹏程系列大模型的发展进行展望；第5节对本文进行总结。

1 可持续学习相关研究

1.1 可持续学习方法概述

可持续学习方法旨在解决模型学习过程中的灾难性遗忘（catastrophic forgetting）问题^[15]，即通过新任务学习新知识时会遗忘旧任务中学到的知识。针对灾难性遗忘问题，可持续学习方法按照不同策略对遗忘问题进行处理，并按照策略类型将可持续学习法划分为三大类^[16]，分别为数据回顾法、参数正则法和参数独立法。

数据回顾法指保留一部分旧数据继续参与训练。旧数据通过直接选取（排演法）、生成法、约束优化法等进行选择。例如，排演法在训练新任务的同时，显式地对有限的存储样本子集进行重新训练。典型的研究有类增量学习器iCaRL^[17]，它存储每个类的样本子集，在学习的特征空间中选择最佳近似类均值；持续原型进化法CoPE^[18]将最近均值分类器法与基于存储层的采样法结合，或者给定随机输入的先前的模型的输出用于近似先前的任务样本^[19]。生成法利用模型的生成能力，不直接存储样本，而是训练一个可以生成之前任务的训练样本的

网络来模拟生成原始数据中学到的样本内容,并结合新数据重新进行学习^[20]。然而,排演法容易过度拟合存储样本的子集,并且似乎受到了联合训练的限制。约束优化法是一种替代解决方案,为后向/前向转移留出更多余地。如 GEM^[21]模型的关键思想是只限制新任务更新而不干扰以前的任务, A-GEM^[22]模型将问题泛化,从之前任务数据缓冲区中随机选择样本,并对样本估计的一个方向进行投影, Aljundi R 等人^[23]提出将 A-GEM 模型中的方法扩展到没有任务边界的纯在线持续学习中,选择最大限度地逼近历史数据可行区域的样本子集。

参数正则法相比数据回顾法避免了存储原始输入,无须优先考虑策略,并减少了内存需求。该方法不引入原数据,而是在损失函数中引入参数正则项,通过对参数正则约束来保留原模型中学习到的知识。Silver D L 等人^[24]首先提出在给定新任务输入的情况下使用之前的模型输出(如把之前任务的输出作为软标签使用),主要是为了提高新任务的性能。Rannen A 等人^[25]在相应的低维空间中学习约束任务特征。类似的方法还有 EWC^[26]和 VCL^[27]等,估计模型参数的分布,并将参数分布作为先验信息,帮助模型在新数据中进行学习。

参数独立法采用独立的模型参数来学习新数据或者新任务,不影响原有训练好的模型。训练完后与原模型结合或者每个任务对应一个独立的模型。通常,此类模型在新任务学习期间会通过参数级或者单元级操作隐藏原有任务模块中参数的更新。不过,这类方法受限于多头设置,无法应对任务之间的共享。Expert Gate 方法^[28]通过自编码门控机制控制不同任务之间的学习共享。参数独立法将不同任务对应的网络参数进行分离,为每个任务设置其专属的数据流通路,能够在一定程度上解决灾难性遗忘问题,然而任务间的结合性比较弱。

近几年,一些其他领域的学习策略也在推动可持续学习方法的发展^[29],主要包括迁移学习、多任务学习、自学习、课程学习、提示学习等。与上述可持续学习方法不同,这些方法侧重于推动模型在特定任务上学习能力的提升。特别是这些学习方法与大模型的结合,促进大模型拥有持续学习的能力。例如,ERNIE 2.0^[30]预训练过程可以分为两个步骤,即构建无监督预训练任务和通过多任务学习增量地更新 ERNIE 模型。这里不同的任务被有序

地加入框架中训练,框架通过持续多任务学习使得模型在学习新任务时不遗忘已经学到的知识。K-Adapter^[31]是一个能够保持预训练模型原始参数不变并支持通用知识注入模型的框架,其对每种注入的知识都使用一个神经适配器进行存储。不同适配器之间没有信息流,因此可以以分布式方式有效地训练多个适配器。此外,还可以借助混合专家 MoE^[32]将多个模型合并为一个模型,为不同的任务或输入选择不同的子模型进行处理,实现模型结构的持续扩展。

上述可持续学习方法以及其他的学习策略为大模型的可持续学习提供了技术基础,然而大模型自身的特点(结构复杂、参数量大、信息存储丰富)决定现有方法很难直接应用在大模型上进行持续学习,以下挑战仍然是当前大模型正在面临且亟须解决的。

① 大模型对不同任务的协调学习能力有待提升,预训练完成的大模型能够在特定任务上通过微调范式取得很好的效果,然而模型却忽略了不同任务间的协调性、相辅性,从而限制了不同任务互补带来的能力提升。在现有的学习策略中,迁移学习只关注后续任务的学习而忽略了前置任务的效果,单纯的多任务学习并未衡量任务之间的关联性,如何让大模型通过不同任务进行协调学习是本文尝试解决的第一个问题。

② 特定类型大模型结构无法动态调整,如多语大模型无法动态地增加新语言的翻译功能。多语大模型包含多种语言的翻译,然而却不支持对模型中未包含的语言与其他语言的翻译。现有工作中混合专家模型可以实现模型结构的动态调整,然而依然不支持翻译语言的扩增,因此如何让多语大模型实现翻译语言的动态调整是本文尝试解决的第二个问题。

对预训练大模型与可持续学习法结合的研究仍处于探索阶段,除了以上两点还面临着其他的挑战,这主要是因为大模型的结构复杂、参数量大、信息存储丰富等。如何从任务、数据、结构等方面对大模型进行改进,使其具有可持续学习能力是目前研究的重点内容。

1.2 鹏程系列大模型的可持续学习能力技术体系

鹏程系列大模型结合实际应用从任务、数据、结构、知识等方面对大模型展开研究,探讨模型多样化的持续学习能力,鹏程系列大模型的可持续学习技术应用生态如图1所示。

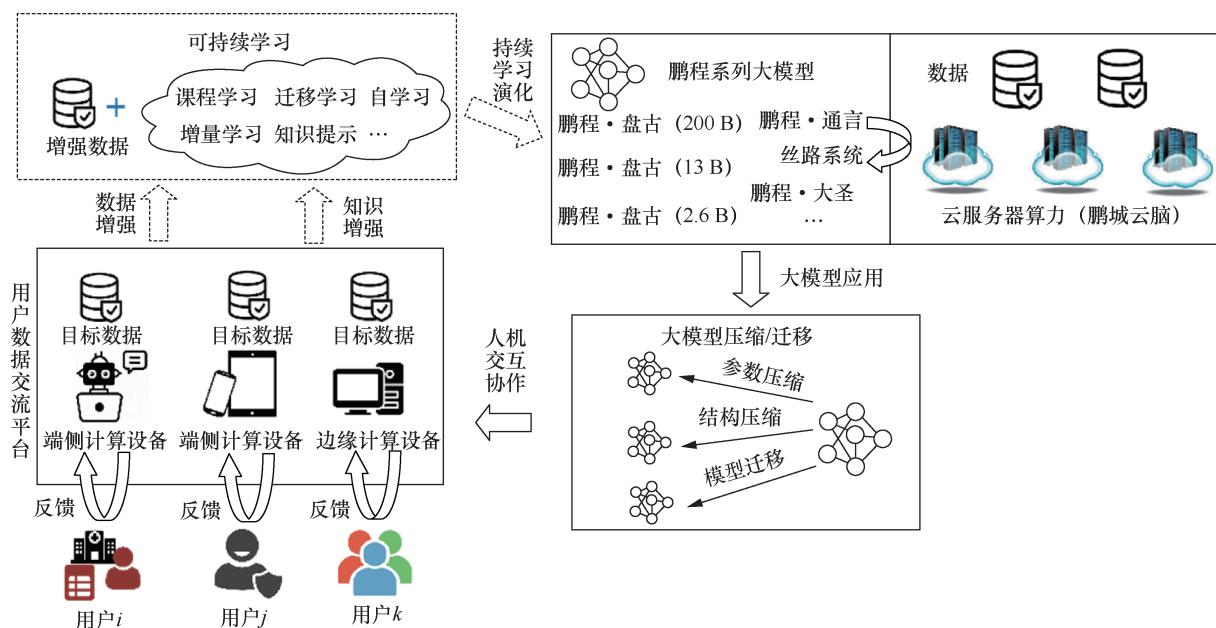


图1 鹏程系列大模型的可持续学习技术应用生态

首先,在用户数据交流平台上收集用户使用过程中产生的数据,平台支持各类端侧设备的接口,可以面向不同用户进行目标数据收集汇总。其次,平台对目标数据、知识进行整理,并结合持续学习方法,如课程学习、迁移学习、增量学习等,将数据传入鹏程系列大模型中进行模型持续学习演化。鹏程系列大模型依托鹏城云脑云服务器算力以及离线收集的大规模多任务数据进行预训练,包含鹏程·盘古系列生成模型,鹏程·通言、丝路系统等。鹏程系列大模型在用户数据上持续学习后,通过模型压缩与迁移技术转化为平台级应用,重新在用户数据交流平台进行应用,演进后的模型可以改善用户的使用体验。在鹏程系列大模型的可持续学习技术应用生态重复此数据收集、模型演进、模型应用过程,实现大模型的应用生态。

1.2.1 基于任务扩展的大模型持续学习技术

在预训练大模型基础上开展任务的可扩展学习,如特定任务预训练、多任务辅助预训练等,是鹏程系列大模型的一个技术研究方向。

例如,在与机器翻译相关的扩展任务中,多语言翻译的目标函数是不同语言间的文本匹配(输入语言与输出语言的匹配),与现有的预训练语言模型目标函数(掩码单词预测)相差较大。本文围绕语义对齐技术进行多语可持续预训练任务学习,语义对齐包括词对齐、短语对齐、句内文本段对齐以及整句对齐,分别从词语、短语、短文本等不同粒度持续进行多语的语义对齐预训练任务构建,该预

训练任务的训练目标函数与多语任务的目标函数类似,从不同粒度进行多语间的语义匹配。最后再通过整句对齐训练(即多语翻译语料)对模型进行完整的多语匹配学习。

1.2.2 基于数据增量融入的大模型持续学习技术

将回顾法和课程学习结合进行数据的可持续增量学习是鹏程系列大模型的另一个技术研究方向。

在回顾法中,新数据和部分原数据混合对预训练模型进行可持续增量训练,结合课程学习法对原数据进行由简到难的动态筛选过程。课程学习是一种训练策略,在选择训练数据时按照先简单后困难的方式依次进行训练,模拟人的学习过程^[33]。初始训练时课程学习方法对原数据进行难度划分并进行难易选择。在训练过程中,课程学习通过先后学习成果对比,对遗忘知识进行数据补充。训练过程重复此过程,则能够持续地从新数据中学习,又可以保证从原数据中学到的信息不被过度遗忘。

1.2.3 基于知识推理的大模型持续学习技术

通过实体/关系向量嵌入、知识实体/关系预测预训练任务对预训练模型进行可持续的知识推理学习,是鹏程系列大模型第三个技术研究方向。

在原有预训练模型基础上,对鹏程系列大模型进行知识微调学习。先对原始预训练语料进行知识实体对齐,原始语句作为输入,将与原预训练模型的输出层向量对齐的知识词元进行反词元化操作,获取潜在知识实体表示,同时向知识实体表示输入关系提示或者实体提示进行关系/实体预测。知识提

示学习任务通过预训练任务将知识推理信息融入原模型的输出表示中，完成知识推理的融入过程^[34]。同时，新知识也可以通过提示任务持续地进行融入学习，不会对模型整体结构造成影响。

2 鹏程·盘古多任务可持续学习的实践探索

持续学习关注单一模型在持续而来的任务上的顺序表现。当前持续学习的研究主要集中在非预训练模型或小模型，受限于非预训练小参数模型本身的能力和容量，这不可避免地会导致灾难性遗忘^[15]，即训练好当前任务后在之前任务上的性能会变差。相对而言，大规模预训练语言模型具有两大优势：巨大的模型容量、通过预训练获得的强大的知识储备和表示能力。巨大的模型容量在理论上可以处理更多的任务，有利于提升模型对旧任务的记忆能力，强大的知识储备和表示能力有利于模型学习新的任务。基于此，本节研究了预训练大模型的多任务持续学习，基于预训练大模型的优势利用持续学习的方式动态增加预训练大模型的任务处理能力。

2.1 基于预训练大模型的多任务持续学习方法

基于鹏程·盘古开源大模型的多任务持续学习拥有大规模预训练语言模型本身的两大优势。本文还创新性地从3个方面利用大规模预训练语言模型的特性来提升其持续学习的性能，基于鹏程·盘古大模型的多任务持续学习如图2所示。其一，基于提示（prompt）将所有类型的任务范式转换成生成任务范式以消除下游任务与预训练任务之间的鸿沟，将下游任务与预训练任务统一起来，无须再为每个任务设计微调方法以适配下游任务，无须为每个任务增加额外的参数空间（为每个任务分配单独的输出层），充分利用大模型预训练阶段获得的能力，降低下游任务间的差异性，促进任务间的知识迁移。其二，利用回顾

法与参数正则法结合的方式进一步提升模型的抗遗忘能力。其三，消除数据量差异影响，传统的回顾法是按固定的数据量进行回顾的，本文优化为按照不同回顾任务中数据量的比例进行回顾。

2.2 实验设置

本文采用任务增量学习模式^[15]，已知一系列任务，模型一次只接收一个任务的训练数据进行训练，直到收敛，符合持续扩展预训练模型（如鹏程·盘古^[2]）的能力要求。

2.2.1 任务与数据

为了兼顾广泛性和通用性，本文选取了6个包含不同任务的数据集，覆盖自然语言理解和自然语言生成类任务。对于每个任务，如果能获取到测试集则在测试集上进行评估，否则在验证集上进行评估。为了减少计算资源消耗，部分任务从数据中随机采样部分子集进行评估。下面分别对每个任务和数据集进行介绍，不同任务中数据统计信息见表1，表1中ACC表示准确率，EM（exact match）表示精确匹配准确率。

（1）完形填空任务，相应的数据集为CMRC（Chinese machine reading comprehension）2017^[35]。CMRC2017包括两个子任务，本文仅采用完形填空子任务。

（2）阅读理解任务，相应的数据集为CMRC2018^[36]。CMRC2018是片段抽取式的阅读理解任务，即从给定的段落中抽取一个片段来回答问题。

（3）自然语言推理（natural language inference, NLI）任务，相应的数据集为CMNLI^[36]。NLI任务需要模型识别两个句子之间的关系：蕴含、中立或矛盾。

（4）文本匹配任务，相应的数据集为AFQMC^[36]。文本匹配任务需要模型识别句子对的相似性：相似、不相似。

（5）实体识别任务，相应的数据集为CLUENER^[37]。

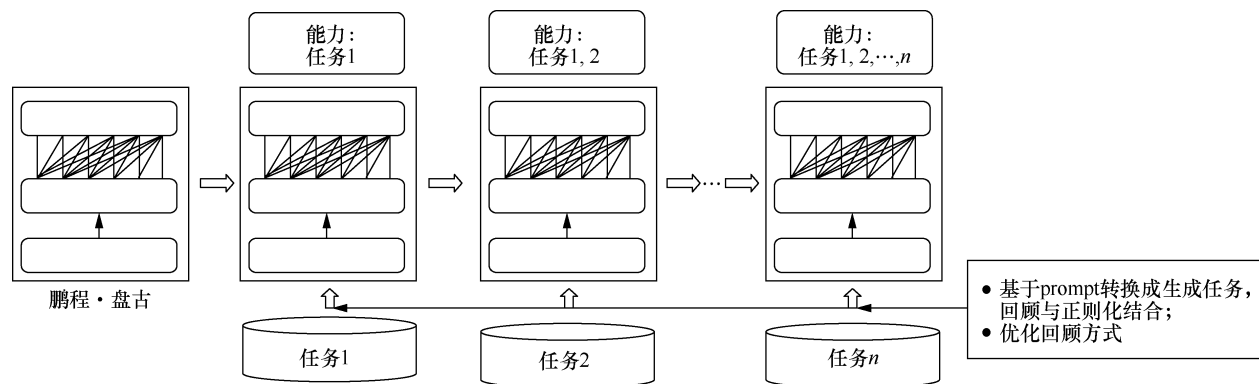


图2 基于鹏程·盘古大模型的多任务持续学习

实体识别任务需要模型识别文本中指定类型的实体。

(6) 文本分类任务, 相应的数据集为 TNEWS^[36]。文本分类任务需要模型从给定识别空间中识别文本的类别, TNEWS 是一个包含 15 个类别的新闻标题分类数据集, 本文实验中为每个样本随机采样 3 个负样本组成一个 4 分类任务以降低计算量。

2.2.2 实验配置

基于 prompt 将所有不同任务范式转换为生成任务范式以消除下游任务与预训练任务的鸿沟, 将下游任务与预训练任务统一起来。具体而言, 本文为每个任务设计至少 3 种 prompt, 不同任务的 prompt 示例见表 2, 表 2 中“{text}”和“{question}”分别表示原始的输入和问题, “{options}”表示候选项, “{ans}”表示待模型生成的结果。

由于每个任务的数据质量相差较大, 本文对每个数据集做了统一的数据预处理。主要的数据预处理包括异常字符处理、连续标点与换行、截断等。预处理后的数据样本被嵌入设计好的 prompt 模板中便形成了最终训练数据。基于快速验证和数据平衡的考虑, 本文限制了每个任务最终训练数据的最大样本量为 50 000 个。

通过回顾与参数正则法结合的方式进一步提升

模型的抗遗忘能力。将回顾方式优化为按回顾任务的数据量的比例。参数正则法采用较大的参数权重衰减系数, 在切换任务前后约束模型的重要参数尽量不发生剧烈的变化。此外, 还采用了权重舍弃法(dropout)等正则化手段。因为所有的任务都转换成了文本生成任务, 所以采用与预训练阶段相同的训练目标, 即下一词预测的交叉熵损失。其他主要实验配置见表 3。

2.3 实验结果

与大多数持续学习方法的评测指标一样^[15], 本文主要从两个方面衡量模型最终的性能, 即平均精度(average accuracy)和平均遗忘率(average forgetting)。平均精度代表模型学习知识的能力, 平均遗忘率代表模型对知识的遗忘程度。在实验中, 本文定义 $a_{i,j}$ 为模型在任务 $i(\text{task}_i)$ 上训练完后在任务 $j(\text{task}_j)$ 上的评估结果, 并定义 T 为任务的总量, 具体指标定义如下。

平均精度 A_i 代表模型训练完 task_i 后整体性能的平均精度, 如式(1)所示:

$$A_i = \frac{1}{i} \sum_{j=1}^i a_{i,j} \quad (1)$$

平均遗忘率 F_i 表示模型训练完当前任务后对之

表 1 不同任务中数据统计信息

任务	数据集	评测指标	评估数据	训练集	验证集	测试集
完形填空	CMRC2017	EM	测试集	300 000	2 000	3 000
阅读理解	CMRC2018	EM	验证集	10 142	1 002	3 219
NLI	CMNLI	ACC	验证集	391 782	12 426	13 880
文本匹配	AFQMC	ACC	验证集	34 334	4 316	3 861
实体识别	CLUENER	ACC	验证集	10 748	1 343	1 345
文本分类	TNEWS	ACC	验证集	53 360	10 000	10 000

表 2 不同任务的 prompt 示例

任务	数据集	prompt
完形填空	CMRC2017	填空: \n{text}\n{question} {ans}
阅读理解	CMRC2018	阅读理解: \n{text}\n 问: {question}\n 答: {ans}
NLI	CMNLI	关系判断: \n{text1}\n{text2}\n 选项: {options}\n 答案: {ans}
文本匹配	AFQMC	语义相似度判断: \n{text1}\n{text2}\n 选项: {options}\n 答案: {ans}
实体识别	CLUENER	实体识别: \n{text}\n{entity_type}: {ans}
文本分类	TNEWS	文本分类: \n{text}\n 选项: {options}\n 答案: {ans}

表 3 其他主要实验配置

基础模型	参数量	新增参数	权重衰减	dropout	优化器	学习率	回顾比例
鹏程·盘古(2.6 B)	26 亿	0	0.1	0.1	Adam	$10^{-6} \sim 10^{-5}$	1%~5%

前任务的遗忘程度，如式（2）所示：

$$F_i = \frac{1}{i-1} \sum_{j=1}^{i-1} f_{i,j}, f_{i,j} = \max(a_{i,j}) - a_{i,j}, j \in \{1, 2, \dots, i-1\} \quad (2)$$

为了进行对比，本文测试了每个任务独立训练时能获取的性能，以及将所有任务进行混合多任务训练时的性能，可以将这两种方式的结果看作持续学习的最佳上界。盘古模型的多任务持续学习性能测试结果见表4，盘古模型的多任务持续学习各训练阶段的平均遗忘率如图3所示。

表4中“原始”表示盘古模型在任务上的原始精度，“混合”表示将所有任务进行混合多任务训练得到的精度，“独立”表示每个任务独立训练时能获取的精度，“持续”表示采用本文持续学习方案训练完所有任务后的精度。实验结果表明，各种学习策略相比原始性能都有了极大的提升，证明了盘古模型强大的任务适应能力。混合和独立训练策略的性能基本持平，任务间的差异性较大，但平均精度却仅相差0.005，证明所选任务能比较公平地测试模型的持续学习性能。与回顾持续学习相比，独立训练平均精度下降1.955，仅有2.95%相对精度损失，验证了基于prompt将任务转换成生成任务并结合回顾和参数正则法能有效地对盘古模型进行持续学习。

表4 盘古模型的多任务持续学习性能测试结果

任务	原始	混合	独立	持续
CMRC2017	37.83	63.2	62.033	58.566
CMRC2018	1.21	46.039	47.934	43.46
CMNLI	50.2	71.96	71.72	68.96
AFQMC	59.29	71.825	71.362	70.25
CLUENER	0	60.522	60.567	60.748
TNEWS	60.95	84.04	83.94	83.84
平均	34.913	66.264	66.259	64.304

图3中，横坐标表示持续学习的各训练阶段，纵坐标表示各个阶段的平均遗忘率。实验结果表明，各个阶段的平均遗忘率均小于5%，最终处于2%水平，并且随着训练阶段的进行有下降趋势，表现出了极好的抗遗忘性能。这验证了基于prompt将任务转换成生成任务并结合回顾和参数正则法的盘古模型多任务持续学习具有优秀的抗遗忘性性能。

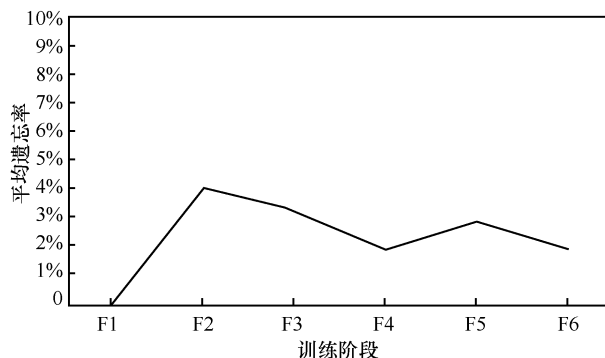


图3 盘古模型的多任务持续学习各训练阶段的平均遗忘率

3 鹏程·通言大模型的可持续学习能力探索

作为人工智能的一个重要分支，机器翻译的目的是突破人类沟通的语言障碍。在世界仍在使用的6 000余种语言中，仅数百种语言可在互联网上广泛采集到并用于模型训练。语言分布不均是有问题的多语言机器翻译模型面临的重大问题^[38]。高资源语料与低资源语料样本存在数量级的差异，导致模型在低资源语料上的学习难度较高。除此之外，不同语言之间的先天差异性也会导致机器翻译学习难度的不均衡。一般来说，即使使用相同规模的训练数据，同一语系或语族下的语言翻译比不同语系或语族下的语言翻译更容易。这导致与语言共性相对较高的语言相比，模型对语言共性相对较低的语言的学习能力较弱。因此，多语言机器翻译本质上是不平衡的，处理这种不平衡对于推进多语言机器翻译至关重要。相比于基于枢轴语言的机器翻译方法（如任意两种语言之间以英文为中枢语言转译），多语言机器翻译系统除了具有运行、训练、维护便捷的优势，还能在稀疏资源的语言翻译方向上带来性能增益。本文在多语言机器翻译上做了学术和工程探索，使用云脑设备训练出132亿参数的多语言机器翻译——鹏程·通言大模型，该模型可以支持多语言的任意方向互译，并在翻译性能上具有大幅提升。

鹏程·通言大模型是在脸书开源的M2M-100 12亿参数模型结构上进行改进的多语言机器翻译模型，通过参数复用和增量式混合专家技术进行多语言联合训练，将模型参数从12亿提升至132亿，在多个小语种语言上的翻译性能大幅提升。同时，鹏程·通言132亿参数大模型仅需一张32 GB显存的V100显卡即可推理，且鹏程·通言的推理速度与12亿参数量模型保持相当。使用动态回顾的终身学习方法，鹏程·通言大模型还可持续学习新的翻译语言。

3.1 参数结构的增量式学习

模型参数的增量式学习是指继承一部分训练好的模型参数，同时增加模型参数的总量。Net2Net^[39]提出将低维度的多层感知机（multilayer perceptron, MLP）的参数通过合理地复制和拆分“等效”地转化为一个高维度的全连接层，这里“等效”是指在相同输入数据下模型的输出不变。维度是指多层感知机的隐藏层节点数，将模型从低维“等效”地扩充为高维（可以在后续训练中继承低维模型的性能），加快收敛速度和提升泛化能力^[40]。

混合专家模型是在条件计算^[41]概念上演化出的一种计算路径，是根据输入样本而变化的神经网络结构模型。在条件计算中，部分网络的活动以样本为基础，因此在混合专家模型中通常存在专家层与门控层。混合专家原理如图4所示，所有专家层并列存在，并且在一次前向传播中并非全部参与运

算，通常由门控层根据输入样本决定哪个专家层参与计算。以 Transformer 结构为例，通常将多层感知机层设置为专家层，将额外的全连接层作为门控层，门控层的输出维度等于专家层的数量。相比于相等参数量的稠密模型，混合专家模型（又称稀疏模型）的优势是其训练的收敛速度和推理速度更快，可以简单地通过增加专家层的数量提升模型的拟合能力，而仅需控制激活专家的数量恒定即可保持模型恒定的训练与推理速度。

不同于简单地增加模型参数量，在鹏程·通言大模型的预训练中，研发人员首次提出转化稠密模型部分结构为混合专家模型的专家层进行增量式训练，相比于从头训练一个翻译模型，增量式训练可以节约 10 倍以上的计算成本（如电力、训练时间等）。增量式混合专家原理如图5所示，增量式训练方法采用将已训练完成的模型和分布式混合专

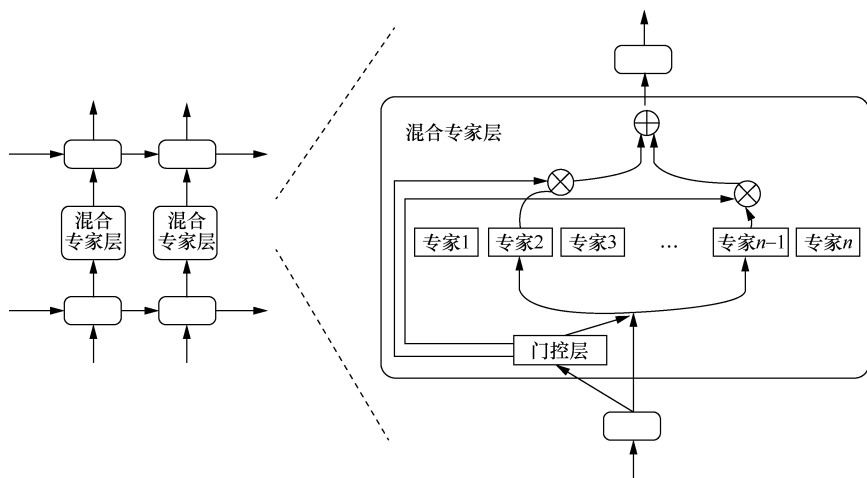


图4 混合专家原理

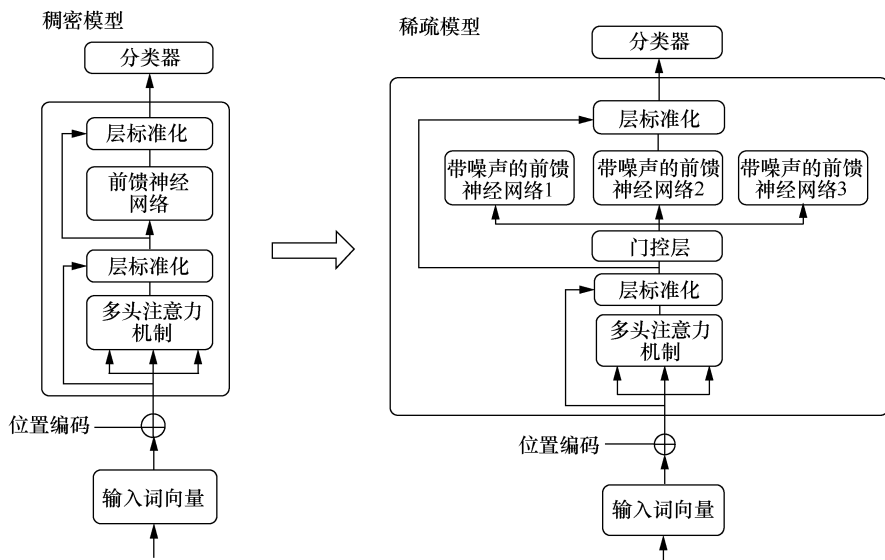


图5 增量式混合专家原理

家结合的策略，通过引入随机噪声将原始模型扩大为包含多个差异化的专家层的神经网络，虽然扩大了模型规模，但并不提升对硬件的容量需求。增量式预训练方法可以适用于社区开源的各种语言模型，具有灵活和可扩展的技术优势。将稠密模型的MLP层参数加入随机噪音作为混合专家模型的不同专家层初始化值，得到的混合专家模型在训练时具有更高的训练起点和收敛速度。两种方式的收敛速度如图6所示。图6描述了从头训练一个混合专家模型和使用本文提出的增量式混合专家训练策略下负对数困惑度的变化曲线，从实验结果可以看出，使用增量式方法可以带来更高的训练起点和更快的收敛速度。

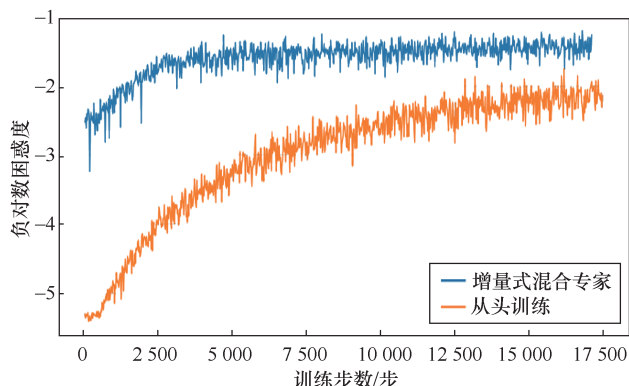


图6 两种方式的收敛速度

鹏程·通言大模型还设计了高效的推理机制，经过对多路分布式专家模块进行专家间的集成整合与稀疏计算，该服务目前已开放集成在丝路机器翻译系统中。虽然鹏程·通言 132 亿参数模型与 M2M-100 12 亿参数模型的推理速度几乎一致，但是使用相同数据训练时，鹏程·通言 132 亿模型在 17 个小语种语言翻译方向的平均 BLEU (bilingual evaluation understudy) 值大幅超越了 M2M-100 12 亿参数模型。

3.2 翻译语言的可持续学习

可持续性学习的主要目标就是在计算和存储资源有限的条件下，在学习新数据的同时尽可能地解决灾难性遗忘问题。语言的可持续学习是指一个多语言机器翻译系统可以持续地学习新的语言翻译方向。语言的持续增加是一个典型的持续学习场景，本文希望多语言模型能够处理持续增加的语言，并且在学习新的语言时能够利用已学知识的信息，加速对新语言的学习。对于学习过的语言，不需要重复的训练就可以保持甚至提升性能。但是深度神经网络模型在持续学

习中通常会面临严重的遗忘问题，如何在学习新语言的同时保持旧语言的翻译能力是多语言翻译模型持续学习的研究方向。

LsED^[42]提出使用一种语言可持续学习的多语言机器翻译系统，该系统通过组合语言特殊的编码器和解码器实现任意方向的语言互译，学习一种新语言仅需要增加一对编码器-解码器，使用 N 对编码器-解码器就可以实现 $(N \times (N-1))/2$ 个翻译方向。不同于 LsED 增量式模型结构的方法，鹏程·通言的设计建立在超大的多语言词表基础上，从词表来看，鹏程·通言已经可以支持世界上主流的百种语言。通过超大且覆盖度全面的词表设计，鹏程·通言仅需要在固定的模型结构上考虑语言的可持续学习问题。

在考虑如何持续学习时，鹏程·通言提出了基于多语言模型编码采样的动态回顾可持续学习机制，模型在持续地学习新语言的同时仍能保持对旧语言的翻译能力，极大程度地解决了神经网络在持续学习时面临的灾难性遗忘问题。动态回顾方法通常将旧任务数据按照一定比例存储在内存中，在学习新任务时可以通过不同形式传递训练信号给当前模型，使模型学习新任务时仍然能够以一定频率复习旧的知识。该方法虽简单有效，但缺陷是模型效果对内存大小以及回顾频率过于敏感，并且回顾的数据被模型学习次数较多，具有更高的过拟合风险。本文通过研究动态的内存更新机制与采样策略，设计更优的学习信号和优化策略提高内存中的样本的训练效率，以减小动态内存的大小，而不是单纯地重复训练旧的数据。通过使用 K -means 方法对同一语言的训练数据的文本向量编码进行聚类，将靠近聚类中心的样本作为“代表性”学习样本，以一定频率回放少量此类样本就可以较好地缓解遗忘，同时还能降低内存占用和训练开销。在这个过程中，编码器的参数是随着训练更新的，需要以一定的频率更新翻译文本的编码和聚类结果，同时内存中的样本也会随之动态地更新。在持续学习方案中，鹏程·通言还实现了基于伪代码生成的回顾学习机制，可以在原始数据不可见的情况下进行增量式训练，为保证原始数据隐私安全提供技术手段。

目前鹏程·通言单个模型支持中文和 17 种稀疏资源语言的任意（共 306 个）方向互译，并支持新语言的可持续学习。相比于基准模型，鹏程·通言在小语种语言-中文方向的 BLEU 值平均提升 4.70，中文-小语种语言方向的 BLEU 值平均提升

表5 部分语言的 BLEU 值

语言	Transformer (多模型)		M2M-100 (官方版)		M2M-1.2B (微调)		鹏程·通言 (单模型)			
	中文-xx	xx-中文	中文-xx	xx-中文	中文-xx	xx-中文	中文-xx	提升	xx-中文	提升
意大利语	17.10	27.87	15.00	31.12	26.10	32.64	29.90	12.80	38.67	10.8
捷克语	14.88	31.50	10.90	29.18	16.50	32.90	18.20	3.32	35.43	3.93
荷兰语	15.21	27.46	14.20	32.92	21.40	37.44	24.20	8.99	42.13	14.67
葡萄牙语	17.46	27.79	15.80	32.99	27.40	36.89	28.40	10.94	40.84	13.05
印度尼西亚语	13.06	18.77	15.60	31.14	25.10	35.29	27.20	14.14	38.89	20.12
希伯来语	16.21	16.89	8.20	12.23	16.10	18.51	18.00	1.79	20.32	3.43
波斯尼亚语	11.38	16.23	4.40	11.80	11.90	17.46	13.50	2.12	19.39	3.16
希腊语	8.20	13.98	6.50	11.49	14.80	16.81	17.10	8.90	18.72	4.74
克罗地亚语	14.32	16.86	5.20	12.77	14.60	17.74	15.70	1.38	19.98	3.12

3.02, 总体性能平均提升 19%, 部分语言的 BLEU 值见表 5。

鹏程·通言的可持续学习方案可以将旧语言的遗忘率控制在 1%~2%, 同时对新语言的学习效果能与预期 (相比基准线) 相当或小幅提升。

4 展望

目前, 鹏程系列大模型针对任务扩展、数据增强、网络参数结构调整以及知识推理能力等方面开展大模型的可持续学习研究, 并取得了初步的成果。然而, 本文涉及的研究仍然处于起步阶段, 特别是这些方面的研究并不是一蹴而就的, 还有大量的工作需要完善, 同时大模型的持续学习还有其他未解决的问题亟须研究。

(1) 大模型的词表更新问题: 目前的大模型词表中尽量涵盖了较多的词汇, 然而大模型在新数据学习过程中, 可能引入原模型中未包含的词汇, 针对该问题, 需要研究如何将新的词汇更新到原有词表中。

(2) 大模型学习的语义表示空间在特定任务上的映射问题: 大模型在预训练任务上学习到语义表示空间, 然而在特定任务上持续学习过程中, 可能面临语义空间不一致的问题, 针对该问题, 需要研究如何向特定的语义空间进行映射。

(3) 大模型持续学习新数据时与原模型中知识冲突问题: 大模型从数据中学习新知识, 但是新知识与原模型中存在的知识可能产生事实冲突, 例如时间先后发生的事实变化、人物关系变化等, 针对该问题, 需要研究如何解决新旧数据中的知识冲突。

大模型可持续学习过程中存在的诸多问题证明了该研究的必要性以及紧迫性, 在相应研究过程中还会面临更具体的细节问题, 目前大模型持续学习研究尚处于起步阶段, 不断地总结旧问题、发现新问题是推动该研究的基础和先决条件。

5 结束语

本文主要介绍鹏程系列大模型在持续学习方面的能力探索, 首先介绍了鹏程系列大模型结合实际情况在任务扩展、数据增强融入、知识推理等方面开展的可持续学习技术体系, 接着介绍了鹏程·盘古开源大模型通过多任务持续学习法对大模型可持续学习能力进行实践探索, 然后介绍了鹏程·通言大模型通过参数复用和增量式混合专家技术进行多语言联合训练的持续学习能力, 最后结合现有研究总结大模型持续学习过程中依然存在的若干挑战, 为今后的鹏程系列大模型的可持续学习提供了参考方向。文章涉及的模型与技术将坚持开源路线, 通过开源共享实现技术持续创新, 逐步形成大模型与可持续学习等关键技术的开源生态。

致谢

感谢鹏程·盘古团队和鹏程·通言团队为本文提供的技术支持。

参考文献:

- [1] 唐杰, 刘洋, 刘知远, 等. 认知与可持续学习: 预训练模型的技术展望[J]. 中国计算机学会通讯, 2021, 17(5): 1-8.
TANG J, LIU Y, LIU Z Y, et al. Cognition and sustainable learning:

- technical prospects of pre-training models [J]. Communications of the CCF, 2021, 17(5): 1-8.
- [2] ZENG W, REN X Z, SU T, et al. PanGu- α : large-scale autoregressive pretrained Chinese language models with auto-parallel computation[J]. arXiv preprint, 2021, arXiv:2104.12369.
 - [3] DEVLIN J, CHANG M W, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding[C]//Proceedings of 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg: Association for Computational Linguistics, 2019: 4171-4186.
 - [4] LAMPLE G, CONNEAU A. Cross-lingual language model pretraining[J]. Advances in Neural Information Processing Systems, 2019, 32: 7059-7069.
 - [5] LIU Y H, OTT M, GOYAL N, et al. RoBERTa: a robustly optimized BERT pretraining approach[J]. arXiv preprint, 2019, arXiv:1907.11692.
 - [6] XUE L T, CONSTANT N, ROBERTS A, et al. MT5: a massively multilingual pre-trained text-to-text transformer[C]//Proceedings of 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg: Association for Computational Linguistics, 2021: 483-498.
 - [7] BROWN T B, MANN B, RYDER N, et al. Language models are few-shot learners[J]. arXiv preprint, 2020, arXiv:2005.14165.
 - [8] FEDUS W, ZOPH B, SHAZEER N. Switch transformers: scaling to trillion parameter models with simple and efficient sparsity[J]. arXiv preprint, 2021, arXiv:2101.03961.
 - [9] YANG A, LIN J Y, MEN R, et al. Exploring sparse expert models and beyond[J]. arXiv preprint, 2021, arXiv:2105.15082.
 - [10] PETERS M, NEUMANN M, IYYER M, et al. Deep contextualized word representations[C]//Proceedings of 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg: Association for Computational Linguistics, 2018: 2227-2237.
 - [11] SONG K, TAN X, QIN T, et al. MASS: masked sequence to sequence pre-training for language generation[C]//Proceedings of the International Conference on Machine Learning. [S.l.:s.n.], 2019: 5926-5936.
 - [12] LIU Y H, GU J T, GOYAL N, et al. Multilingual denoising pre-training for neural machine translation[J]. Transactions of the Association for Computational Linguistics, 2020, 8: 726-742.
 - [13] CHEN Z Y, MA N Z, LIU B. Lifelong learning for sentiment classification[C]//Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers). Stroudsburg: Association for Computational Linguistics, 2015: 750-756.
 - [14] MERMILOD M, BUGAISKI A, BONIN P. The stability-plasticity dilemma: investigating the continuum from catastrophic forgetting to age-limited learning effects[J]. Frontiers in Psychology, 2013, 4: 504.
 - [15] MAI Z D, LI R W, JEONG J, et al. Online continual learning in image classification: an empirical survey[J]. Neurocomputing, 2022, 469: 28-51.
 - [16] DELANGE M, ALJUNDI R, MASANA M, et al. A continual learning survey: defying forgetting in classification tasks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021: 1.
 - [17] REBUFFI S A, KOLESNIKOV A, SPERL G, et al. iCARL: incremental classifier and representation learning[C]//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2017: 5533-5542.
 - [18] DE LANGE M, TUYTELAARS T. Continual prototype evolution: learning online from non-stationary data streams[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. [S.l.:s.n.], 2021: 8250-8259.
 - [19] ROBINS A. Catastrophic forgetting, rehearsal and pseudorehearsal[J]. Connection Science, 1995, 7(2): 123-146.
 - [20] GOODFELLOW I, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial nets[J]. Advances in Neural Information Processing Systems, 2014, 27: 2672-2680.
 - [21] LOPEZ-PAZ D, RANZATO M A. Gradient episodic memory for continual learning[J]. Advances in Neural Information Processing Systems, 2017, 30: 6467-6476.
 - [22] CHAUDHRY A, RANZATO M, ROHRBACH M, et al. Efficient lifelong learning with A-GEM[C]//Proceedings of the International Conference on Learning Representations. [S.l.:s.n.], 2019.
 - [23] ALJUNDI R, LIN M, GOUJAUD B, et al. Online continual learning with no task boundaries[J]. arXiv preprint, 2019, arXiv:1903.08671.
 - [24] SILVER D L, MERCER R E. The task rehearsal method of life-long learning: overcoming impoverished data[M]//Advances in artificial intelligence. Heidelberg: Springer Berlin Heidelberg, 2002: 90-101.
 - [25] RANNEN A, ALJUNDI R, BLASCHKO M B, et al. Encoder based lifelong learning[C]//Proceedings of 2017 IEEE International Conference on Computer Vision. Piscataway: IEEE Press, 2017: 1329-1337.
 - [26] KIRKPATRICK J, PASCANU R, RABINOWITZ N, et al. Overcoming catastrophic forgetting in neural networks[J]. Proceedings of the National Academy of Sciences of the United States of America, 2017, 114(13): 3521-3526.
 - [27] NGUYEN C V, LI Y Z, BUI T D, et al. Variational continual learning[J]. arXiv preprint, 2017, arXiv:1710.10628.
 - [28] ALJUNDI R, CHAKRAVARTY P, TUYTELAARS T. Expert Gate: lifelong learning with a network of experts[C]//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2017: 7120-7129.
 - [29] BIESIALSKA M, BIESIALSKA K, COSTA-JUSSÀ M R. Continual lifelong learning in natural language processing: a survey[C]//Proceedings of the 28th International Conference on Computational Linguistics. PA: International Committee on Computational Linguistics, 2020: 6523-6541.
 - [30] SUN Y, WANG S H, LI Y K, et al. ERNIE 2.0: a continual pre-training framework for language understanding[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(5): 8968-8975.
 - [31] WANG R Z, TANG D Y, DUAN N, et al. K-Adapter: infusing knowledge into pre-trained models with adapters[C]//Proceedings of the Findings of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2021.
 - [32] MA J Q, ZHAO Z, YI X Y, et al. Modeling task relationships in multi-task learning with multi-gate mixture-of-experts[C]//Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. New York: ACM Press, 2018: 1930-1939.
 - [33] BENGIO Y, LOURADOUR J, COLLOBERT R, et al. Curriculum learning[C]//Proceedings of the 26th Annual International Conference on Machine Learning. New York: ACM Press, 2009: 41-48.
 - [34] 社会芳, 王昊奋, 史英慧, 等. 知识图谱多跳问答推理研究进展、挑战与展望[J]. 大数据, 2021, 7(3): 60-79.

DU H F, WANG H F, SHI Y H, et al. Progress, challenges and research trends of reasoning in multi-hop knowledge graph based question answering[J]. Big Data Research, 2021, 7(3): 60-79.

- [35] CUI Y, LIU T, CHEN Z, et al. Dataset for the first evaluation on Chinese machine reading comprehension[C]//Proceedings of the 11th International Conference on Language Resources and Evaluation. [S.l.:s.n.], 2018: 2721-2725.

- [36] XU L, HU H, ZHANG X W, et al. CLUE: a Chinese language understanding evaluation benchmark[C]//Proceedings of the 28th International Conference on Computational Linguistics. PA: International Committee on Computational Linguistics, 2020: 4762-4772.

- [37] XU L, DONG Q Q, YU C, et al. CLUENER2020: fine-grained name entity recognition for Chinese[J]. arXiv preprint, 2020, arXiv: 2001.04351.

- [38] 亢晓勉, 宗成庆. 融合篇章结构位置编码的神经机器翻译[J]. 智能科学与技术学报, 2020, 2(2): 144-152.

KANG X M, ZONG C Q. Fusion of discourse structural position encoding for neural machine translation[J]. Chinese Journal of Intelligent Science and Technology, 2020, 2(2): 144-152.

- [39] CHEN T Q, GOODFELLOW I, SHLENS J. Net2Net: accelerating learning via knowledge transfer[J]. arXiv preprint, 2015, arXiv: 1511.05641.

- [40] KAPLAN J, MCCANDLISH S, HENIGHAN T, et al. Scaling laws for neural language models[J]. arXiv preprint, 2020, arXiv:2001.08361.

- [41] SHZAEER N, MIRHOSEINI A, MAZIARZ K, et al. Outrageously large neural networks: the sparsely-gated mixture-of-experts layer[J]. arXiv preprint, 2017, arXiv:1701.06538.

- [42] ESCOLANO C, COSTA-JUSSÀ M R, FONOLLOSA J A R, et al. Multilingual machine translation: closing the gap between shared and language-specific encoder-decoders[C]//Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2021: 944-948.

[作者简介]



余跃(1988-), 男, 博士, 国防科技大学副研究员, 鹏城实验室网络智能部开源所技术总师, 主要研究方向为智能化软件工程、群体智能、开源生态等。



刘欣(1989-), 男, 博士, 鹏城实验室网络智能部开源所博士后, 主要研究方向为自然语言处理领域大模型持续演进、语义匹配、词义学习、胶囊网络应用等。



蒋芳清(1989-), 男, 鹏城实验室网络智能部开源所工程师, 主要研究方向为自然语言处理、预训练大模型、小样本学习、持续学习等。



张晗(1993-), 男, 鹏城实验室网络智能部开源所联培博士生, 主要研究方向为自然语言处理中的可持续学习、小样本学习、多语言机器翻译、预训练语言模型等。



王晖(1968-), 男, 博士, 鹏城实验室网络智能部开源所研究员, 主要研究方向为自然语言处理、分布式机器学习、联邦学习等。



曾炜(1973-), 男, 博士, 北京大学副研究员, 主要研究方向为计算机体系结构、分布式系统、计算机视觉等。